SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities

Year: 2023 | *Citations: 496* | *Authors: Dong Zhang, Shimin Li, Xin Zhang*

Abstract

Multi-modal large language models are regarded as a crucial step towards Artificial General Intelligence (AGI) and have garnered significant interest with the emergence of ChatGPT. However, current speech-language models typically adopt the cascade paradigm, preventing inter-modal knowledge transfer. In this paper, we propose SpeechGPT, a large language model with intrinsic cross-modal conversational abilities, capable of perceiving and generating multi-model content. With discrete speech representations, we first construct SpeechInstruct, a large-scale cross-modal speech instruction dataset. Additionally, we employ a three-stage training strategy that includes modality-adaptation pre-training, cross-modal instruction fine-tuning, and chain-of-modality instruction fine-tuning. The experimental results demonstrate that SpeechGPT has an impressive capacity to follow multi-modal human instructions and highlight the potential of handling multiple modalities with one model. Demos are shown in https://0nutation.github.io/SpeechGPT.github.io/.